Acceptability-based QoE Models for Mobile Video

Wei Song and Dian W. Tjondronegoro

***Abstract*—Quality of experience (QoE) measures the overall perceived quality of mobile video delivery from subjective user experience and objective system performance. Current QoE computing models have two main limitations: 1) insufficient consideration of the factors influencing QoE, and 2) limited studies on QoE models for acceptability prediction. In this paper, a set of novel acceptability-based QoE models, denoted as *A-QoE*, is proposed based on the results of comprehensive user studies on subjective quality acceptance assessments. The models are able to predict users' acceptability and pleasantness in various mobile video usage scenarios. Statistical nonlinear regression analysis has been used to build the models with a group of influencing factors as independent predictors, including encoding parameters and bitrate, video content characteristics, and mobile device display resolution. The performance of the proposed A-QoE models has been compared with three well-known objective Video Quality Assessment metrics: PSNR, SSIM and VQM. The proposed A-QoE models have high prediction accuracy and usage flexibility. Future user-centred mobile video delivery systems can benefit from applying the proposed QoE-based management to optimize video coding and quality delivery strategies.**

***Index Terms*— Acceptability, mobile video, modeling, pleasantness, quality of experience (QoE).**

I. Introduction

To improve the quality of mobile video services, research from academics and industry service providers have focused on developing quality of experience (QoE) models to predict overall user-perceived quality for optimizing quality provision. However, modeling QoE is challenging due to the complex influences of user experience and diverse conditions of video content, network bandwidth, and mobile devices.

Many objective video quality assessment (VQA) metrics, such as structural similarity (SSIM) [1], multiscale SSIM (MS-SSIM) [2], and NTIA general model of video quality metric (VQM) [3], have been widely used as the QoE models. However, these metrics need reference videos, and their efficiency in predicting overall quality of mobile video has not been fully studied. Reference-free QoE models have used network-related factors such as encoding parameters and video content features as predictors to estimate the users' mean opinion scores (MOS) of received video quality [4-6]. However, it is argued that the MOS-based measurement are unable to indicate whether video quality is acceptable or not [7, 8].

Telecommunication standardization sector of International Telecommunication Union (ITU-T) defines QoE as the end-user's overall acceptability of a service or application [9]. However, a little research has focused on establishing models to predict the user acceptance of mobile video [7, 10-12]. The work could be improved by involving more influencing factors of user experience such as device characteristics and use context. In addition, previous research on user's acceptance threshold may become inadequate in reflecting user experience for pleasant viewing [13].

To address these limitations in QoE modeling of mobile video, this paper has focused on the following objectives:

- Develop QoE models based on user-centered acceptability for pleasant viewing
- Examine the performance of some well-known VQA metrics, used as objective QoE models, for predicting user acceptability

In order to build the dataset required for modeling the QoE, we conducted two user studies that involved a total of 80 participants, two types of mobile devices (iPhone 3GS and iPhone 4), and 870 test clips from 15 video sources. In these studies, participants were asked to select the lowest acceptable and the lowest pleasing quality using a customized mobile app while viewing a set of different groups of video qualities. The lowest acceptable quality means that below this quality, users are not willing to watch; the lowest pleasing quality refers to the quality they feel would be comfortable enough for regular viewing, while being mindful of reducing cost in data consumption.

Binary logistic regression analysis was used to determine the significant factors influencing the user acceptance. Based on the influencing factors and the relationship between lowest acceptable and lowest pleasing quality metrics, we established a set of acceptability-based QoE (A-QoE) models to predict quality acceptability by mapping its relationship with the key influencing factors. These A-QoE models can be used for various purposes, including quality control for mobile video coding, and automatic quality decision for mobile video delivery.

The rest of the paper is structured as follows. Section 2 discusses the related work, and Section 3 describes the details of the user study data collection and analysis. Section 4 discusses the QoE modeling process and the proposed A-QoE

W. Song is with Information Systems School, Science and Engineering Faculty, Queensland University of Technology, Brisbane, 4001 Australia (e-mail: w1.song@ qut.edu.au).

D. W. Tjondronegoro is with Information Systems School, Science and Engineering Faculty, Queensland University of Technology, Brisbane, 4001 Australia (e-mail: dian@qut.edu.au).

models. In Section 5, we evaluated the performance of three full-reference objective VQA metrics, PSNR, SSIM, and VQM, by comparing their correlations with subjective acceptability measures. Discussion and conclusion are given in Section 6 and 7, respectively.

II. Related Work

Modeling QoE is challenging due to the difficulties in representing a complex subjective measure of user experience in a simple and objective way. Generally, QoE models are constructed by three steps: (i) collecting subjective evaluation data; (ii) identifying critical elements (or operations) influencing the subjective value; and (iii) determining the relationship between the subjective value and these elements. [4, 14, 15].

Objective metrics for perceptual video quality assessment (VQA) are often used as objective QoE (oQoE) in video services [5]. These models focus on the impact of low-level video characteristics on human visual system (HVS) and are developed to fit in Mean Opinion Scores (MOS) gained from subjective assessments. Common objective VQA metrics include PSNR, SSIM [1], MS-SSIM [2], and NTIA general model VQM [3]. Although PSNR was not developed based on subjective assessments, the heuristic mapping between PSNR and MOS [16] has been widely used. The performance of these VQA metrics has been evaluated by comparing the correlation between the objective scores and the subjective assessment scores [17]. The results showed that both the MS-SSIM and the VQM metrics performed well, while the PSNR was the worst. However, a study evaluating perceptual quality of scalable video content on mobile screens indicate that the PSNR is slightly better than the SSIM and VQM metrics [18]. These conflicting conclusions warrant further study to investigate how the objective VQA metrics can be used to evaluate the subjective quality of mobile video. Advanced metric such as MOtion-based Video Integrity Evaluation (MOVIE) [19] provides better performance than those common VQA metrics [17]. However, we did not attempt to examine its performance for estimating mobile video quality due to its high computing complexity.

Reference-free QoE models rely on seeking the factors that cause the quality loss in the entire video delivery process. ITU-T has recommended many VQA metrics for quantifying the QoE of an audiovisual service as perceived by the end user [20]. For example, the E-model (Recommendation G.107) [21] predicts the quality affected by various transmission impairments of bandwidth, delay, jitter and loss. The opinion model (Recommendation G.1070) evaluates video quality based on packet loss and coding distortion under the combination of bitrate and frame rate [22]. Recent Additive Log-Logistic Model (ALM) [6] is formulated by better capturing the relationship of visual quality against lossy compression and transmission error (slicing and freezing) and by taking into account the content features of content unpredictability and motion homogeneity to achieve better accuracy. In mobile video streaming scenario, two types of models were proposed in [23] to estimate video quality for the most frequent content types: news, soccer, cartoon, panorama, and rest. One is based on average bitrate and four content characteristics of motion; the other is a content dependent low complexity metric based on bitrate and frame rate for each content class. Major QoE models are established to predict 5 or 11 scales of MOS. However, it is argued that the scales are not sufficient to determine acceptable quality for end users [8]. Binary measure is therefore suggested to be used in assessing acceptability of mobile TV (videos) [24, 25].

Based on ITU-T's definition of QoE as the overall acceptability [9], it should encompass not only user's perception for video quality, but also user's desire and need. One psychological method for measuring people's acceptance is known as Method of Limits, which is often accomplished by asking participants to decide whether or not they accept the quality of various videos viewed in successive discrete steps, as an ascending or descending series [14]. Only a few researchers have worked on QoE modeling based on acceptability. In [10], M2A models are built to measure the extent of a quality being acceptable based on the full-reference metric VQM, thus not suitable for real-time QoE management. In [7, 11], QoE models are proposed for six video content types (news, sports, animation, music, comedy and movie) and three viewing devices (mobile phone, PDA, and laptop), based on a linear combination of bitrate and frame rate. These models do not consider other influencing aspects, such as video resolution, and are dependent on the ability of knowing the video content types. In [12], a decision tree of audiovisual quality acceptance is determined by means of network type, transport protocol, video quality, and user watching behavior. The research made in living context, yet the subjective assessments for only two video qualities may not be enough to delegate the usage situations.

There is a little research that determines the relationships between acceptability and MOS. The G.107 E-model provides mapping formula from MOS to the binary measure of Good or Better (GoB) and Poor or Worse (PoW) [21]. However, some researchers found that the G.107 e-model overestimates the actual acceptability for mobile TV, and they proposed a set of more precise mapping formula M2A for different content types [10]. However, the M2A may not be sufficient for videos with bigger resolution than 320×240 pixels (which was evaluated in [10]) due to a lack of consideration in the effect of image resolution. Another study [8] addressed the mapping of MOS to Acceptability for mobile broadband data services. Based on a series of lab and field experiments, they found a consistent mapping between the binary acceptance and the ordinal MOS ratings across different applications, such as web browsing and file downloads. Nonetheless, an acceptable quality does not necessarily mean that the video is pleasant for regular viewing [13]. It motivates our study to develop QoE models to estimate both the lowest acceptability and the pleasing acceptability.

There are two processes in developing QoE models: i) identifying inputs and the respective features of the model, and ii) mapping the features to a quality index [26]. The feature identification depends on the available data, which is often obtained from subjective quality assessments. The

suitable features for a quality model should be those that are directly related with user perception, which are usually identified through statistical analysis techniques. The mapping process is to find the best-fit quality prediction model. There are many different approaches, including discriminant analysis [14], machine learning classification algorithm – decision tree [11, 12], multiple linear regression [15], and non-linear regression [4, 6]. Discriminant analysis and decision tree are suitable for classifying groups, such as acceptable or unacceptable, whereas linear and non-linear regression are appropriate for calculating an index, such as acceptability and MOS. In this paper, we show a novel process to collect user acceptance data through mobile phones, and adopt non-linear regression technique to produce mathematical QoE models for acceptability prediction based on the nature of data fit curve.

III. User Study

To develop A-QoE models, the quality acceptability data was derived from two user studies conducted in 2010 and 2011, denoted as *Study1* and *Study2* respectively. The settings of the studies are described in the following sections.

A. Test Tool

The iPhone 3GS and iPhone 4 were used as test equipment in Study1 and Study2 respectively. Both devices have a 3.5-inch screen, supporting different display resolutions. The screen of iPhone 3GS is 480×320-pixel resolution at 163 ppi, and iPhone 4 is 960×640-pixel resolution at 326 ppi (denoted as "retina display").

B. Test Videos

In total, we used 15 high-resolution (>=1280×720 pixels) videos as sources, consisting of standard and real-world datasets that are depicted in Fig. 1. The seven standard videos [27, 28] are uncompressed YUV 4:2:0 format and include nature scenes and crowd. The eight real-world videos are compressed videos at high bitrates (>3500kbps), 2-4 minutes long, covering five typical content genres of mobile videos: news, movie, music, sports and animation [29-31]. These real-world videos were from recorded broadcast news and soccer matches, movie trailers, a movie segment, music videos, and open movie source [32]. All of the 15 sources (/contents) were used in Study2, while only five of them, "Planet51", "Backupplan", "Miley", "Tennews" and "Sports", were used in Study1.

To produce the test video clips, the video sources were encoded into H.264/AVC format with a set of combinations of encoding parameters: spatial resolution (SR), frame rate (FR) and quantization parameter (QP). The encoding parameters used in Study1 and Study2 are listed in Table I and Table II respectively, of which a little bit of difference is related to the display capability of the test equipment. It should be noted that in Study2 the FR of 12.5fps was applied for only nine contents (i.e., "Planet51", "Backupplan", "Miley", "Tennews", "Sports", "Oldtown", "Parkjoy", "Pedestrain" and "Shields") due to the consideration of assessment time and necessity. Eventually, a total of 870 degraded video sequences were generated as the test videos, where 150 sequences (2FR×3SR×5QP×5Content) were used in Study1, and 720 sequences with 450 (3SR×10QP×15Content) encoded at 25fps and 270 (3SR×10QP×9Content) encoded at 12.5fps were used in Study2.

For evaluation purpose (details in subsequent section D), these produced test videos were assembled in groups of 10 for each content, shown in column 1 of Table I and II. In each quality group, the 10 video clips are arranged in a bitrate order. Study2 used a finer QP level than Study1 in order to make the participants feel that the quality-change transition smoother.

C. Participants

A total of 80 people were recruited, 40 in Study1 and 50 in Study2, with 10 of them taking part in both. There was a gender balance (20 male, 20 female) in Study1, and 27 females and 23 males in Study2. These participants have different ages (between 17 and 40 with an average of 26.24), experiences of viewing videos on mobile phones, and study/career backgrounds (including education, marketing, information, administration, and nursing).

In Study2, 35 participants were involved in the assessment for the 450 test videos encoded at 25fps of frame rate, and 15 were involved into the evaluation of the 270 test videos at 12.5fps.

D. Procedure

We designed the subjective assessment process as a scenario-based evaluation task, and guided through a customized iPhone application, depicted in Fig. 2. The participants were allowed to adjust the video quality within a quality group while they were watching. Their task was to select the lowest acceptable quality and the lowest pleasing quality from each video quality group. The lowest acceptable quality refers to the quality below which one is not willing to watch; the lowest pleasing quality refers to the quality that one feels good enough for regular and comfortable watch.

Using the test application, after a participant randomly chose one of video contents (Fig. 2a) to watch, the video played starting from the lowest or the highest quality within one test video group. Swiping left or right on the screen could adjust the video quality to be higher or lower gradually within the same group (Fig. 2c). Double tapping on the screen and clicking the relative confirmation button from a pop-up message window (Fig. 2d) could confirm the current video as the lowest acceptable/pleasing one. Once both the lowest acceptable and the lowest pleasing qualities were determined, a "Next" button would appear to allow the participant to evaluate next group. The participant did not need to watch the rest qualities in the same group; however, he/she could change his/her decisions before clicking the "Next" group button. The iPhone application automatically recorded the participants' decisions and stored into the device (Fig. 2b). During the process, a test video was playing in a loop mode. When switching the quality, the next quality of the same video content would start to play from the break point of the content (allowing up to 1-second overlap).

In Study1, the participant was asked to select only the

lowest pleasing quality for each of the 15 test video groups (5 content ×3group). The testing time was about 20 minutes. In Study2, the participant was required to choose both the lowest acceptable and the lowest pleasing quality. It took around 30-45 minutes to complete the 270 videos (encoded at 12.5fps), and around one hour for the 450 videos (encoded at 25fps). To avoid fatigue, we gave 10-minute break when a participant completed a half of his/her task during the data collection period.

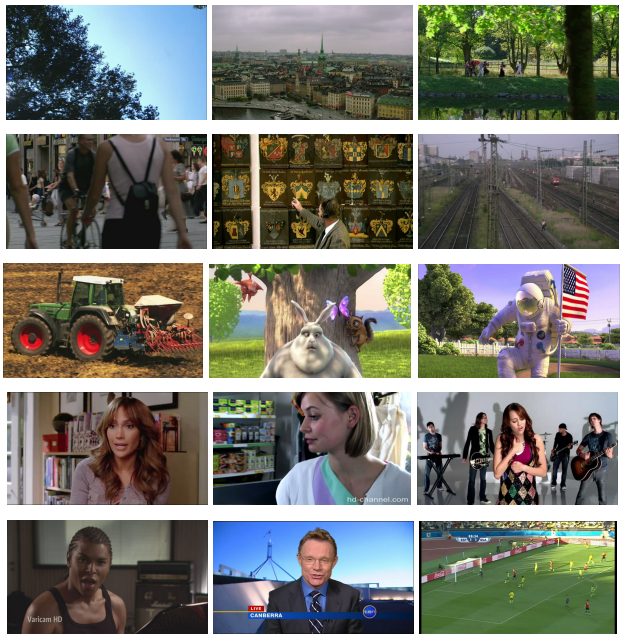


Fig. 1. Thumbnails of video sources. In left-to-right and top-to-down order: Standard: Bluesky, Oldtown, Parkjoy, Pedestrian, Shields, Station2, Tractor; Real: Bigbunny, Planet51, Backupplan, Lucid, Miley, Mountaintop, Tennews, and Sports.

TABLE I
COMBINATION OF ENCODING PARAMETERS FOR 1ST STUDY

Group	SR (pixels)	FR (fps)	QP				
1	320×240	12.5	40	36	32	28	24
	320×240	25	40	36	32	28	24
2	480×320	12.5	40	36	32	28	24
	480×320	25	40	36	32	28	24
3	640×480	12.5	40	36	32	28	24
	640×480	25	40	36	32	28	24

SR = Spatial Resolution, FR= Frame Rate, QP = Quantization Parameter

TABLE II
COMBINATION OF ENCODING PARAMETERS FOR 2ND STUDY

Group	SR (pixels)	FR (fps)	QP									
1	480×270	12.5	40	38	36	34	32	30	28	26	24	22
2	640×360	12.5	40	38	36	34	32	30	28	26	24	22
3	960×540	12.5	40	38	36	34	32	30	28	26	24	22
4	480×270	25	40	38	36	34	32	30	28	26	24	22
5	640×360	25	40	38	36	34	32	30	28	26	24	22
6	960×540	25	40	38	36	34	32	30	28	26	24	22

Fig. 2 The test iPhone application. (a) a list of video content with a demo video has been demonstrated to the participant; (b) an example of recorded data; (c) an instruction of how to use the application; (d) a screenshot of determining the selection of video quality

E. Data Processing

The original assessment data was the subjects' acceptable quality levels, which is then transformed into binary data to denote whether a certain video quality is acceptable/pleasing or not. For each of the participant's records, video clips with lower quality (i.e. less bitrate) than the selected lowest acceptable quality within the same SR and FR group were regarded as "unacceptable" and represented with "0", and the others with equal or greater quality were regarded as "acceptable" and represented with "1". The same transformation was made for the pleasing quality evaluation data. After shifting missing data and outliers, a total of 76870 binary assessments (0/1) were obtained for the 870 test clips from the 80 participants. The outlying cases identified by the studentized residuals less than -2 or greater than +2, which was suggested in [33] for binary logistic regression.

Based on the binary data, the degree of user acceptance for each test clip was computed as the ratio of the accumulation of "1"s in the total number of the ratings. There are two acceptability indexes used: general acceptability *GAcc* and pleasant acceptability *PAcc*. Their computing equations are shown in (1) and (2). Strictly speaking, the GAcc score (Q_{GAcc}) means the possibility of a video quality being generally accepted by viewers and the PAcc score (Q_{PAcc}) means the possibility of a video quality making viewers pleasant or comfortable.

$$Q_{GAcc} = \frac{\text{the number of basic acceptable ratings}}{\text{the total number of ratings}} \quad (1)$$

$$Q_{PAcc} = \frac{\text{the number of acceptable ratings for pleasant watch}}{\text{the total number of ratings for pleasant watch}} \quad (2)$$

Ultimately, 720 of Q_{GAcc} scores and 750 of Q_{PAcc} scores were obtained. These scores, which were originally derived from the subjective assessments, reflect the end-users' overall

perceived quality and therefore will be used as the indicators of QoE in Section 4. They will also be used to assess the performance of objective VQA metrics in predicting user acceptability in Section 5.

IV. Acceptability Based QoE Models

From the user studies, two acceptability indicators, Q_{GAcc} and Q_{PAcc}, have been obtained to represent the quality of experience in accordant with user acceptability to a mobile video under general and pleasant viewing circumstances. To establish the acceptability-based QoE (A-QoE) models, this section firstly examines the relationship between Q_{GAcc} and Q_{PAcc} to determine whether different models need to be established. Then, the section presents the detailed process of modeling QoE, including determination of model predictors, modeling criteria, and models forms and coefficients.

A. Relationship between QGAcc and QPAcc

Fig. 3 shows that there is a close cubic relationship between Q_{GAcc} and Q_{PAcc}. Based on the nonlinear regression analysis, their relationship can be represented as the function (3) with the R^2 value of 0.966, which means that the function accounts for about 96.6% of the Q_{GAcc} variability in the dependent variable Q_{PAcc}. Due to their strong correlation, we only need to build QoE models for the pleasant acceptability, as the general acceptability can be deduced from (3). It should be noted that only the Q_{GAcc} greater than 6.7% could be computed through (3). In fact, it is unnecessary to calculate a very low acceptability because there is no point providing such a low quality of videos to users. Another reason for using Q_{PAcc} is that the pleasant acceptability was investigated for both iPhone 3GS and iPhone 4 devices, and therefore it can reflect the impact of the mobile devices.

$$Q_{GAcc} = \begin{cases} 2.805Q_{PAcc} - 3.28Q_{PAcc}^2 + 1.416Q_{PAcc}^3 + 0.067 & 0 \le Q_{PAcc} < 0.9 \\ 1 & 0.9 \le Q_{PAcc} \le 1 \end{cases} \quad (3)$$

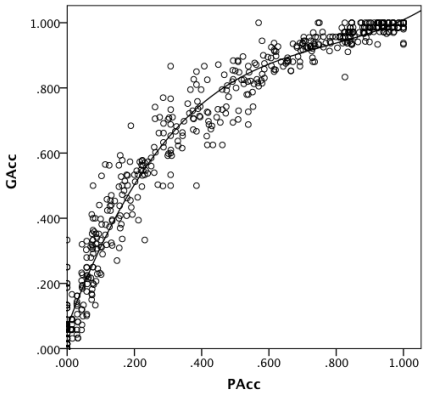


Fig. 3. Relationship between general acceptability and pleasant acceptability of mobile video quality

B. Model Predictors

QoE modeling is the process of establishing the relationship between the QoE indicator (i.e., Q_{PAcc} in this paper) and a series of independent variables (i.e., predictors). Certain factors that significantly affect the subjective quality acceptability were considered as the predictors of our QoE models. To determine these factors, binary logistic regression analysis was conducted based on the original 76870 binary data, where statistical significance level was set as α=0.05. The results revealed a significant impact of QP, SR, FR, bitrate, video content, and mobile device screen resolution ($p<.001$). Therefore, factors associated with these aspects will be considered as potential predictors for the QoE models. To accommodate these aspects into the QoE models, the followings will define the predictors and the normalization computation where required, and discuss their correlations with the acceptability. Table III summarizes the selected variables, which can be used as the possible predictors of a QoE model.

TABLE III
POTENTIAL PREDICTORS IN QOE MODELS

Category	Variable	Description
Displaying device related	SDPPI	Mobile device screen PPI divided by 163
	RVD	Video resolution divided by display device screen resolution
Video coding parameters	SSR	Video resolution divided by 320×240 pixels
	QP	Quantization parameter
	FR	Frame rate
	LBR	Common logarithm of bitrate
CI in sematic definition	CTmovie	Whether a video is a "movie" (1-yes, 0-no)
	CTsport	Whether a video content is about "sport" (1-yes, 0-no)
CI in uncompressed domain	ASI & ATI	Averaged spatial and temporal complexity
	NSI & NTI	Normalized SI and TI
	W	Weight of spatial complexity over temporal complexity
CI in compressed domain	MAI	Mean motion vector (MV) magnitude
	MAD	Mean deviation of MV directions from the dominant direction
	MAP	Proportion of motion in a video

Scaled Device Screen PPI (SDPPI): Different mobile devices have different display features. In this study, the devices used have the same screen size of 3.5-inch but with different screen resolutions. As a result, the mobile devices can be characterized by an index, pixel per inch (PPI), which represents the number of pixels that can be displayed within one inch of a video frame. Our studies have used the iPhone 3GS and the iPhone 4 with two PPI values, 163 and 326 respectively. These PPI values divided by 163 are indicated by a variable *SDPPI* with two values: 1 and 2.

Scaled Spatial Resolution (*SSR*): There is an obvious trend that the acceptability increases with the increase of a video's resolution. However, the video resolution in pixels is too big to be of suitable use in a model, where other predictors have much smaller values. Instead of using it directly, a scaled spatial resolution *SSR* has been adopted. Each SR is scaled by the smallest SR (320×240=76800 pixels) used in this study, thus the resulting SSR values are 1, 1.687, 2, 3, 4, and 6.75 for the respective SR values of 320×240, 480×270, 480×320, 640×360, 640×480, 960×540 pixels.

Ratio of video frame resolution to device resolution (*RVD*): Findings from data analysis showed an interaction video resolution and display device on the acceptability. Fig. 4 illustrates the mean pleasantness of various spatial resolutions

at 25fps for both test devices. It can be observed that PAcc at a given image resolution (e.g., 480×270 pixels) on iPhone 4 is much lower than that of similar resolution (e.g., 480×320 pixels) on iPhone3GS. It indicates that people needs a higher video quality for watching videos on a mobile device with a higher display resolution, more so than if they were using a device with a lower display resolution (this confirms the natural hypothesis). To delegate this correlation, a variable RVD is computed as the value of a video frame resolution divided by the device display resolution (in this case, 480×320 pixels for iPhone3GS and 960×640 pixels for iPhone 4). The resulting RVD values are 0.693, 1.0 and 1.387 for the SR of 320×240, 480×320 and 640×480 pixels on iPhone 3GS respectively, and 0.477, 0.636 and 0.955 for 480×270, 640×360 and 960×540 pixels on iPhone 4 respectively. Fig. 4 shows a reasonable growth of PAcc with RVD increase.

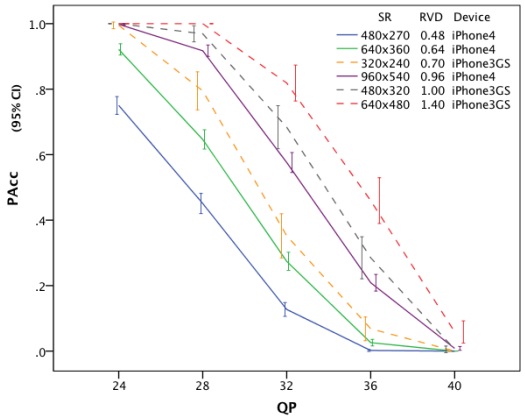


Fig. 4. Impact of QP, SR and RVD on pleasant acceptability

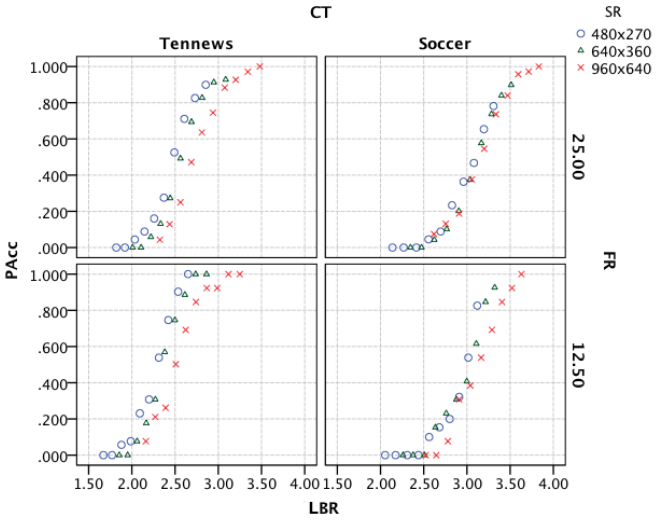


Fig. 5. Impact of LBR and FR on pleasant acceptability for Tennews and Soccer videos

Quantization Parameter (*QP*): From Fig. 4, it can also be observed that the curve of QP to the acceptability has a sigmoidal, or "S" shape, which indicates a logistic relationship. Moreover, the SR or RVD influences the midpoints of the QP_PAcc curves.

Logarithm Bitrate (*LBR*): In our studies, the relationship between QP and bitrate is logarithm at any combinations of SR and FR. To achieve consistency with QP and to avoid using a big value of bitrate, a variable $LBR = \log_{10}(BR)$ has been used. Fig. 5 shows a logistic relationship between LBR and PAcc, which is slightly affected by SR but importantly affected by FR.

Frame Rate (*FR*): Frame rate is closely related to the smoothness of video content and the bitrate reduction, but its impact on user acceptability is highly dependent on the content type. From Fig. 5, it can be seen that for videos with relatively slow motion (e.g., Tennews), a low FR (12.5fps) reaches certain acceptability at a lower bitrate than a high FR (25fps) does; while for videos with relatively fast motion (e.g., Soccer), there is no significant bitrate saving by using a low FR. It may be explained by the fact that frame jump is easier to be perceived when viewing a fast and large movement video with a low FR; and to compensate for the visual distortion, better image quality is required, which leads to the consumption of bitrate that is saved from reducing frame rate. Considering the above discussion, the correlation between FR and video content, and between FR and bitrate need to be considered. The normalization of FR is to divide it by 12.5fps.

Content identification (*CI*): The significant impact of video content on the acceptability was shown in different directions when working with QP and bitrate together. For example, under the control of QP, the acceptability of "movie" video is lower than others such as "music" and "news" videos (based on McNemar tests that compared the subjective assessment 0/1 for each pair of content types, $p<.05$), as people had a higher quality requirement for "movie" videos in order to see human faces and expressions clearly, according to interview data. Under the controlled bitrate, "sports" video has a much lower acceptability than other content types. Fig. 5 shows the comparison between Tennews and Soccer's acceptability at the same LBR, which indicates a higher demand for compressing video content with global and fast motion.

To accommodate video content information as predictors in the QoE modeling, variables representing CI need to be defined. There are a lot of video content features can be used to distinguish videos, but only the features that can generate best fit models will be selected. We have examined a series of CI variables, shown in Table III. Categorical variables *CTmovie* and *CTsport* have been used to denote whether a video belongs to the content type of "Movie" or "Sport" (1-yes, 0-no). These variables somehow reflect user's preference. Content information can be obtained from the description of video sources, but if unavailable, some technical content characteristics need to be extracted. We have applied two approaches to attain the characteristics. The first approach follows the method suggested in ITU-T Recommendation P.910 [34] to calculate spatial information (i.e., complexity) (SI) and temporal information (TI) of a video content. The SI is based on the Sobel filter over the luminance space of a video frame and the TI is based on the temporal difference between successive frames. On the basis of the SI and TI values for each frame, we used the following

variables: *ASI* and *ATI* (the averaged SI and TI), *NSI* and *NTI* (the normalized SI and TI by the maximum values), and *W* (the ratio of the NSI to the NTI, which indicates the relative dominance of spatial complexity over the temporal complexity [35]).

The second approach works in video compression domain for MPEG4 or H.264/AVC formatted videos to extract motion characteristics: *motion activity intensity* (MAI) defined as the mean magnitude of motion vector (MV); *motion activity proportion* (MAP) defined as the proportion of the number of non-zero MVs in the total number of MVs; and *motion activity direction* (MAD) defined as the deviation of MV directions from the dominant movement direction. A high MAI value often indicates fast movement; a big MAP value often indicates large movement areas; a small MAD relates to consistent movement. These motion characteristics have been used to classify content successfully in [23]. The three motion features are calculated by (4).

$$MAI = \frac{1}{N_t}\sum_i \sqrt{x_i^2 + y_i^2}$$
$$MAP = \frac{N_{nz}}{N_t} \qquad (4)$$
$$MAD = \sqrt{\frac{1}{N}\sum_{i=1}^{24} n_i (D_i - D_m)^2}$$

where x and y are the coordination of a motion vector (MV) in H.264 video coding; N_t is the total number of MVs, and N_{nz} is the number of non-zero MVs. Three steps are needed to compute the MAD for a particular frame. First, the whole coordinate is divided by 15 degree into 24 subsectors ($D_i, i = 1,2,...24$), and each MV direction ($\arctan(y/x)$) is mapped into one of the subsectors. n_i is the number of MVs in D_i . Then, the dominant direction D_m is selected where most MV directions belong. Lastly, the deviation of other directions from the dominant direction is accumulated.

C. Curve Fitting

The next important QoE modeling step is to map the relationship between the QoE indicator and the various predictors. The ultimate choice of predictors depends on the curve fit and the usage situations of QoE models. We adopted statistical technique to find the QoE models that can generate the best-fitting estimate of the true acceptability curves. In a QoE model, some variables are unnecessary when their effects are reflected by other variables/correlations, or when they cannot significantly improve the ability of the prediction model. Only the most efficient predictors that have the highest simple correlation for the desired outcome were chosen.

As discussed earlier, there are sigmoidal curves between QP and PAcc and between LBR and PAcc. Thus, a logistic curve model will do a good job to fit the curving sigmoidal shape of the acceptability data. A widely used logistic function is the four-parameter logistic (4PL) [22, 36]. The 4PL model can fit curves in logit-log space well, but cannot effectively model asymmetric data [37]. With our acceptability data, the curves were not symmetrical, that is, the upper curvature and the lower curvature are different, as illustrated in Fig. 6. Therefore, a more suitable function – five-parameter logistic (5PL) model – was chosen for the asymmetric curves [37]. The general formula of 5PL is given by (5), where c>0 and g>0, and setting g=1 leads to the 4PL function. The effects of parameters c, b and g are illustrated in Fig. 6.

$$y = a + \frac{d}{\left(1 + \left(\frac{x}{c}\right)^b\right)^g} \qquad (5)$$

where:
a: estimated value for the minimum asymptote
b: slope factor
c: mid-range concentration
d: estimated value for the maximum asymptote
g: asymmetry factor

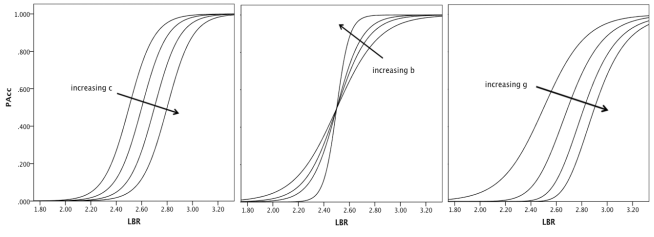


Fig. 6. Asymmetric logistic curve illustrating effects of parameters of a 5PL function. In this example a=0 and d=1.

According to statistical regression theory, a common way to determine a best-fitting model is to find the parameters that minimizes the *sum of squared errors,* also called *residual sum of squares* (RSS). The quality of the curve fit is often accessed by the *R square* (R^2). The R^2 equals the ratio of the regression sum of squares to the total sum of squares (shown in (6)), which explains the proportion of variance accounted for in the dependent variable by the model. by the model. The R^2 has a value between 0 and 1. A value of the R^2 close to 1 means a good curve fit.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation (Total sum of square)}} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \qquad (6)$$

Another way of evaluating the model performance is to examine the correlation between the predicted responses and the observed responses. The *Pearson linear correlation coefficient* (*PC r*) is a measure of the strength of linear dependence between two variables, is used to indicate model accuracy. The *Spearman rank order correlation coefficient* (*SROC rho*) indicates prediction monotonicity. The coefficient r or *rho* is between 0 and 1, and the value equal to ±1 indicates a perfect relationship.

In addition, *root-mean-square error* (*RMSE*) is frequently used to measure the differences between values predicted by a model and the values actually observed. The lower the RMSE is, the more accurate the model is.

D. A-QoE Models

The strong relationship between general acceptability

(GAcc) and pleasant acceptability (PAcc) suggests that we only need to develop models for the PAcc. Therefore, based on the logistic relationships between QP and PAcc and between LBR and PAcc, we propose the following two metrics (7) and (8), containing QP and LBR, respectively. Equation (7) and (8) are in the form of 5PL (refer to (5)) with the minimum asymptote as zero (i.e., *a=0* in (5)) and the maximum asymptote as one (i.e., *d=1* in (5)).

When determining the parameters, C1, B1, and G1 in (7) and C2, B2 and G2 in (8), the factors influencing the changes of the curves (see also Fig. 6) have been considered. For the model using QP as the main predictor, C1 is affected by spatial resolution (SR), displaying device (DEV), frame rate (FR) and video content types or characteristics (CI), and B1 is affected by SR and DEV. For the model using LBR as the main predictor, C2 is affected by SR, FR and CI, B2 is affected by FR and DEV, and G2 is affected by DEV. As a result, equation (7) and (8) become (9) and (10), respectively. These parameter functions are linear combinations of various dependent variables that associate with the influencing aspects, such as SSR and RVD (representing SR), FR, DPPI (representing DEV), CTmovie, W, and MVP (representing CI), and their interrelations. The linear parameter functions are decided due to two main reasons:

- When determining a mathematical function, the linear function is firstly considered due to its simplicity, which is important in a real application.
- We have examined a variety of non-linear combinations (e.g., combinations of SR/FR and content features, quadratic and logarithm functions) to test if a better fit can be obtained. The result were either no improvement or worse than the linear function.

$$Q_{PAcc}(QP)=\frac{1}{\left(1+\left(\frac{QP}{C1}\right)^{B1}\right)^{G1}} \tag{7}$$

$$Q_{PAcc}(LBR)=\frac{1}{\left(1+\left(\frac{LBR}{C2}\right)^{B2}\right)^{G2}} \tag{8}$$

$$Q_{PAcc}(QP)=\frac{1}{\left(1+\left(\frac{QP}{f(SR,FR,DEV,CI)}\right)^{f(SR,DEV)}\right)^{G1}} \tag{9}$$

$$Q_{PAcc}(LBR)=\frac{1}{\left(1+\left(\frac{LBR}{f(SR,FR,CI)}\right)^{f(FR,DEV)}\right)^{f(DEV)}} \tag{10}$$

The functions (9) and (10) can be expressed in diverse manners depending on the different application conditions. We have considered two conditions: (i) whether or not the display features of the mobile device are known; and (ii) which type of video content identification (in Table III) is known. In line with the combinations of these two conditions, several scenarios and their corresponding QoE prediction models (11) – (16) have been developed based on (9) or (10), summarized in Table IV.

TABLE IV
SCENARIOS AND THE CORRESPONDING MODELS

	Aware of device display feature	Unaware of device display feature
Aware of semantic CI	(11)	(12)
Aware of uncompressed domain CI	(13)	(14)
Aware of compressed domain CI	(15)	(16)

The variables in (11) – (16) were determined by stepwise regression to ensure they are statistically independent and significant at the criterion level p<.001 based on the t-tests. The weighting coefficient of each variable in each model, shown in Table V, was determined by nonlinear regression for the overall Q_{PAcc} to minimize the sum of squared error. The model performance is indicated by the values of R^2, RMSE, PC r and SROC rho. Fig. 7 (a-d) show the scatter plots between the subjective PAcc values and the predicted values by QP-based model (11) (13) and LBR-based model (15) and (16) respectively.

$$\frac{1}{\left(1+\left(\frac{QP}{a+b\cdot RVD+(c+d\cdot FR)\cdot CTsport+e\cdot CTmovie}\right)^{h\cdot RVD+i}\right)^{g}} \tag{11}$$

$$\frac{1}{\left(1+\left(\frac{QP}{a+b\cdot SSR+(c+d\cdot FR)\cdot CTsport+e\cdot CTmovie}\right)^{h\cdot SSR+i}\right)^{g}} \tag{12}$$

$$\frac{1}{\left(1+\left(\frac{QP}{a+b\cdot RVD+d\cdot FR\cdot ATI+e\cdot NSI+f\cdot W}\right)^{h\cdot RVD+i}\right)^{g}} \tag{13}$$

$$\frac{1}{\left(1+\left(\frac{QP}{a+b\cdot SSR+d\cdot FR\cdot ATI+e\cdot NSI+f\cdot W}\right)^{h\cdot SSR+i}\right)^{g}} \tag{14}$$

$$\frac{1}{\left(1+\left(\frac{LBR}{a+b\cdot SSR+c\cdot FR+d\cdot MAP+e\cdot MAD}\right)^{h\cdot FR\cdot SDPPI+i}\right)^{g+f\cdot SDPPI}} \tag{15}$$

$$\frac{1}{\left(1+\left(\frac{LBR}{a+b\cdot SSR+c\cdot FR+d\cdot MAP+e\cdot MAD}\right)^{h\cdot FR+i}\right)^{g}} \tag{16}$$

As these models are concerning some conditions of information availability, they are flexible for various applications and use case scenarios. When the information of the mobile device targeted is known, the models involving the display resolution variable (e.g., (11), (13) and (15)) can be used for QoE prediction. When predicting the acceptability of a compressed video, the models (15) and (16) can be utilized. When encoding a video to a targeted acceptability, the models

(11) or (12) can be simply used if content type information (e.g., a sport video) is aware; otherwise (13) or (14) can be adopted with the detected content complexity information.

In general, taking account the mobile screen feature into the QoE models provides more accurate prediction, which can be seen by comparing the performance of (11), (13) and (15) with (12), (14) and (16), respectively. The important influence of device screen resolution on prediction accuracy can also be visualized through Fig. 7(c) and 7(d), whereby the blue 'x' markers indicate the data derived on 163ppi screen (Study1) and the red 'o' markers on 326ppi screen (Study2). According to Fig. 7(d), without considering the screen resolution the video acceptability may be underestimated when viewing on the screen with a small display resolution.

These series of A-QoE models reflect the comprehensive impact of various factors on the user pleasantness. In (11) to (14), because the mid-range concentration (C1) of the QP-PAcc curve increases with the increase of video resolution related parameters (RVD or SSR), the coefficient *b* is positive; because the effect of FR on the QP-PAcc curve is more sensitive for fast "sport" videos, $FR \cdot CTsport$ or $FR \cdot ATI$ are involved to reflect their interrelation; and because "movie" and "sport" had a lower acceptability than other types of videos, the coefficients *c* for *CTsport* and *e* for *CTmovie* are negative in (11) and (12). Furthermore, as a larger resolution (RVD or SSR) leads to a slightly steeper QP-PAcc curve for the same video, the coefficient *h* in the slope function (B1) is positive.

In (15) and (16), the impact of video content is best represented by the motion features MAP and MAD in compressed domain. The two variables indicate how big the movement area is and how complex the movement is. The impact of SSR and FR on the mid-range concentration (C2) of the LBR-PAcc curve is positive. Differing from the QP models, the slope of the LBR-PAcc curve is more sensitive to FR, or the interrelation of FR and DEV. Moreover, the coefficient *f* in (15) reflects the effect of the device feature on the curve's asymmetry.

TABLE V
PARAMETERS FOR A-QOE MODELS AND MODEL PERFORMANCE

Coeffi-cients	Values in the equation					
	(11)	(12)	(13)	(14)	(15)	(16)
a	26.916	33.806	19.89	28.165	1.916	1.901
b	9.951	0.694	9.832	0.673	0.03	0.043
c	-6.455	-5.588	—	—	0.099	0.155
d	3.154	2.717	0.054	0.064	1.174	1.333
e	-2.638	-2.689	5.264	1.947	0.053	0.049
f	—	—	0.549	0.678	1.05	—
g	2.028	2.844	2.370	3.030	-0.313	0.841
h	5.575	0.429	5.867	0.452	2.732	3.454
i	5.913	7.342	5.360	7.120	-25.473	-24.734
R square	0.955	0.900	0.951	0.895	0.929	0.889
RMSE	0.079	0.117	0.082	0.121	0.098	0.124
PC r	0.977	0.949	0.975	0.946	0.964	0.943
SROC rho	0.973	0.948	0.974	0.945	0.955	0.934

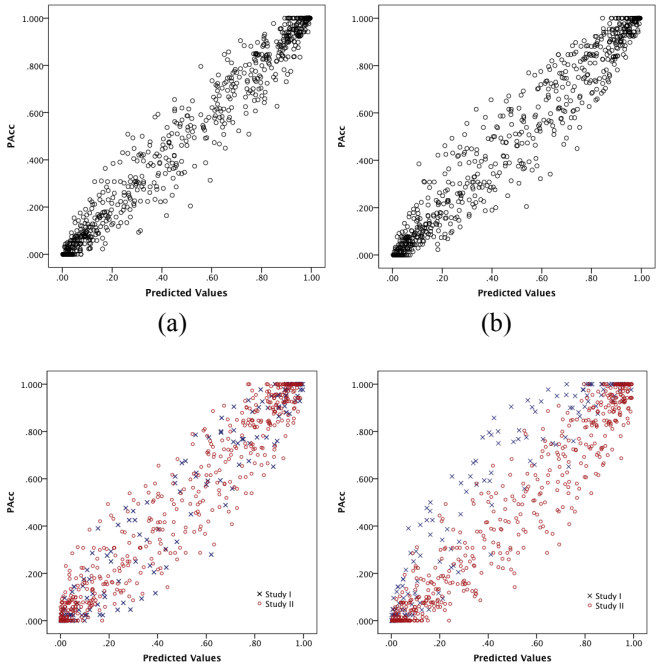


Fig. 7. Scatter plots of predicted acceptability versus observed acceptability by (a) QP Model (11), (b) QP Model (13), (c) BR Model (15), and (d) BR Model (16).

V. OBJECTIVE VIDEO QUALITY ASSESSMENTS VERSUS SUBJECTIVE ACCEPTABILITY

To determine whether the existing VQA models can predict user acceptability, we investigated the performance of three well-known VQA metrics, including PSNR, SSIM, and VQM. For each VQA metric, the overall index of a video is computed by averaging all frame quality scores. Here, the PSNR and SSIM indexes were computed only for the luminance channel (Y component) of the video frame, and directly acquired during the encoding process of the test videos using FFmpeg (http://ffmpg.org). The VQM is based on the NTIA general model, which has been standardized by ANSI and included into two ITU recommendations ITU-I J.144 [38] and ITU-R BT.1683 [39]. The VQM implementation was downloaded from http://www.its.bldrdoc.gov/vqm/. When calculating each test clip's quality score by the VQM metric, the video source was converted into the same resolution and frame rate of the test clip using FFmpeg, but keeping the original quantization quality using the FFmpeg option '–sameq'.

A. VQA Assessments vs. Subjective Acceptability

The accuracy of PSNR, SSIM and VQM metrics in estimating the quality acceptability were evaluated by comparing the correlations between the measured quality values and the acceptability data obtained from the user assessments. To distinguish from the VQA metric's name, the computed quality prediction by each VQA metric will be called as P_xxx, where the 'xxx' is a VQA metric's name, i.e., P_psnr, P_ssim, and P_vqm.

Spearman rank order correlation (SRO) was used to measure the monotonicity between the P_VQA and the QAcc,

and between P_VQA and QPle. Pearson correlation (PC) was used to measure the accuracy of these VQA metrics in predicting the quality acceptability. Due to the nonlinear correlations between the P_VQA values and the subjective scores, the PC was computed after performing a set of nonlinear regression with logistic functions. Table VI shows the performance of the VQA metrics (PSNR, SSIM, and VQM) in terms of SRO and PC coefficients. The logistic relations between these P_VQA values and the QAcc are illustrated in the scatter plots of Fig. 8 (a-c).

The SRO and PC coefficients (rho and r) for the overall video dataset indicate a strong correlation between the VQA metrics of the SSIM and VQM and the quality acceptability, indicated by rho and r >0.6; and a medium correlation between the PSNR and the acceptability, indicated by rho and r 0.5-0.6. This result is consistent with the conclusion from VQA comparison studies [17, 40], where the PSNR has a lower performance than the SSIM and VQM in terms of the accuracy and monotonicity in predicting subjective quality. However, when observing the SROC and PLC coefficients for real and standard video datasets separately, we found some interesting phenomena: the SSIM is generally superior among these metrics (rho and r >0.79); the PSNR has a good performance for real videos (r >0.8), but the worst for standard videos despite of a strong correlation with the acceptability (rho and r >0.6); in contrast to PSNR, the best performance of VQM is manifested in the estimation for standard videos, but underperforms for real videos than the others.

TABLE VI
COMPARISON OF THE PERFORMANCE OF VIDEO QUALITY ASSESSMENT METRICS (PSNR, SSIM AND VQM) IN ESTIMATING QUALITY ACCEPTABILITY

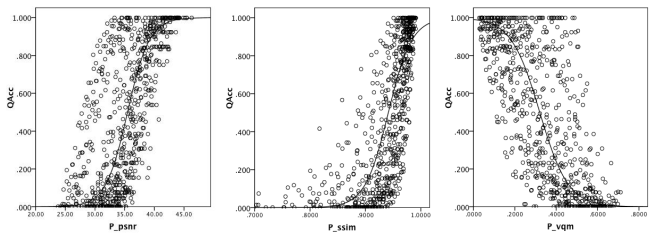
a. Spearman Rank Order Correlation (rho)

Metric	Acceptability (QAcc)			Pleasantness (QPle)		
	All	Real	Std	All	Real	Std
PSNR	0.567	**0.823**	0.638	0.595	**0.827**	0.741
SSIM	**0.745**	**0.869**	**0.819**	**0.754**	**0.824**	**0.851**
VQM	-0.73	-0.74	-0.73	-0.69	-0.65	**-0.87**

b. Pearson Linear Correlation (r)

Metric	Acceptability (QAcc)			Pleasantness (QPle)		
	All	Real	Std	All	Real	Std
PSNR	0.539	**0.834**	0.645	0.588	**0.830**	0.738
SSIM	0.659	**0.857**	**0.798**	**0.707**	0.796	**0.845**
VQM	**-0.686**	-0.723	-0.704	-0.685	-0.634	**-0.867**

Note: Correlation is significant at the 0.01 level (1-tailed); the highlighted are the top value in each column and the values greater than 0.8.



(a) (b) (c)

Fig. 8. Scatter plots of VQA scores versus subjective acceptability for (a) PSNR, (b) SSIM, and (c) VQM.

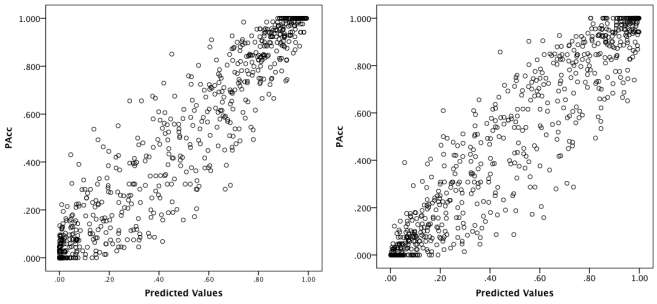
We also examined the factors that mostly affect the correlation between each VQA index and the acceptability. After examining the effect of SR, FR, and content type on the three VQA metrics, we found that SR affected SSIM and VQM more than PSNR. Under a controlled SR, FR mostly affects PSNR and VQM, but not SSIM. With a fixed FR and SR, the standard videos have a looser distribution than the real videos and locate themselves in a different area from the real videos for PSNR and SSIM. This explains why the overall correlation between PSNR and acceptability is low, while the separate correlations for real videos and standard videos are high.

B. VQA Metrics vs. proposed A-QoE Models

According to the above analysis, the objective measures of video quality derived from the three VQA metrics (i.e., PSNR, SSIM and VQM) have a close correlation with the subjective quality acceptability, evidenced by the SROC and PC coefficient greater than 0.6. However, these VQA metrics cannot provide high accuracy of acceptability prediction. The maximum correlation coefficient R is equal to 0.867 (from VQM metric for standard videos in Table VI (b)), which can explain up to 75.2% of pleasantness variation (R^2=.752). Compared to these models, the developed A-QoE models can reach at least 12.5% to 21.3% higher explanation (with the R^2 from 87.8% to 96.5% shown in Table V).

Given that PSNR and SSIM algorithms have a low computational complexity and have been used in rate-distortion control of video coding, it is worthwhile bringing them into the QoE modeling. We have shown how to achieve PAcc prediction models that uses PSNR and SSIM in (18) and (19), which prediction performances are shown in Fig. 9 (a) and (b) respectively. As discussed earlier, PSNR-PAcc curve is sensitive to the changes of SR, FR and content features, and SSIM-PAcc curve is sensitive to SR and content features. Therefore, our models take these effects into consideration. For instance, the coefficient g in (19) varies with the variables for representing video resolution (i.e., *RVD*) and video content (i.e., *NSI*). Moreover, the models involve a new variable *isStd* (0–real-world videos; 1–standard videos) to indicate the more significant impact of content features of standard videos on the PSNR and SSIM values. The prediction performance is given in Table VII. Comparing these PC r and SROC rho values in Table VII to those in Table VI, it can be seen that the improvement of the prediction accuracy is significant (see also the last column of Table VII).

The computation of the VQM has a much higher complexity than PSNR and SSIM, therefore we did not try to use it as the predictor for A-QoE modeling. Furthermore, compared to these full-reference models, the reference-free A-QoE models have more flexibility in usage.



(a) (b)

Fig 9. Scatter plots of predicted acceptability versus observed acceptability by (a) PSNR Model (18) and (b) SSIM Model (19)

$$Q_{PAcc}(PSNR) = \frac{1}{\left(1+\left(\frac{P\_psnr}{49.03-2.17RVD-0.28FR-8.24NSI-0.195ASI\cdot isStd}\right)^{-14.52-2.06W}\right)^{0.067FR}} \quad (18)$$

$$Q_{PAcc}(SSIM) = \frac{1}{\left(1+\left(\frac{P\_ssim}{0.997-0.923RVD-0.016NTI}\right)^{-172.52}\right)^{.069-0.13RVD-0.52NSI-0.025W\cdot isStd}} \quad (19)$$

TABLE VII
PERFORMANCE OF MODEL (18) AND (19)

Metric	Index	All	Real	Std	Overall Improve
PSNR(18)	R square	0.911	0.927	0.874	-
	RMSE	0.111	0.102	0.129	-
	PC r	0.955	0.953	0.935	**0.367**
	SROC rho	0.947	0.964	0.929	**0.352**
SSIM(19)	R square	0.889	0.904	0.857	-
	RMSE	0.124	0.117	0.137	-
	PC r	0.943	0.951	0.926	**0.236**
	SROC rho	0.939	0.946	0.920	**0.185**

.

VI. DISCUSSION

In this paper, we propose user-driven A-QoE models that consider the impact of subjective and objective quality assessments. The A-QoE models offer several contributions to the current attempts for QoE modeling.

The A-QoE models take into consideration significant influencing factors of user acceptability, which were discovered from a comprehensive user study. Thus, they are expected to achieve a closer prediction of user values. User data collection involving both standard and real-world video materials may help to obtain more realistic user values.

The A-QoE models are easy to use as we used a unified expression, as opposed to different models for each type of video contents (e.g [4]) and display devices (e.g [7]), by integrating the most important feature of a mobile device (i.e., screen resolution) and the video content features.

The A-QoE models are novel due to its capability of predicting pleasant-quality for regular viewing, rather than just lowest acceptable to watch. We also revealed the relationship between general acceptability and pleasant acceptability. This will be helpful for video providers to evaluate their service quality from different levels of user acceptance.

The A-QoE models can be utilized in a wide area of applications, based on the available and derivable information. For example, the QP-based models can be used in acceptability-based quality control for mobile video coding through adjusting the encoding parameters; the bitrate-based models can be used to determine an optimal quality of mobile videos based on the network bandwidth, or to predict the acceptability of a given video where its bitrate is known.

The A-QoE models predict the acceptable degree of a quality, in contrast to QoE models that only predict whether a quality is acceptable or not [7]. This provides more flexibility for the video providers, as they can decide to what extent they would like to delight users through providing a certain range of video quality, in order to satisfy the diversity of users' requirements and preferences, mobile devices, and network conditions.

While many studies concentrate on network-parameter (such as packet loss and error rates) for QoE modeling [4, 15], this study primarily focuses on the aspect of video coding and assumes that network transmission is controlled by other strategies. However, we still take an important network issue – network bandwidth – into consideration. The encoding bitrate is related to the requirement of bandwidth. The bitrate-related QoE models can be used to manage which quality is to be delivered to the end user based on the change of network bandwidth.

Regarding to the correlation between the predicted (objective) quality scores by PSNR, SSIM, and VQM and the subjective acceptability scores (Table VI), we can conclude that these full-reference metrics are effective for predicting (subjective) acceptability to some degree. Their prediction performance can be significantly enhanced by considering the influences of video resolution, frame rate, display characteristics, and content features. Comparing the overall performance among the three metrics, SSIM and VQM are better than PSNR, which conform to the study [17], where PSNR's performance in estimating MOS was reported poorly. However, when applied to our real-world video data, simple metric PSNR closely relates to the subjective acceptability. Considering the advantage of its simplicity, it is argued that properly utilizing the PSNR measurements (e.g., combining with FR and content features) is able to provide an acceptability-based quality control in video encoding processing (see also (18)). Similarly, SSIM-based acceptability metric is also useful in quality control (see also (19)).

In addition to the modeling study, a few interesting issues related to video content have been revealed. When observing the scatter plots between the acceptability and QP, LBR, P_psnr, and P_ssim, we found that the real-world videos had a more concentrated distribution than the short-segmented standard videos. This is associated with the high diversity of content characteristics for these standard videos. For a long-duration video, the calculation of a content identification goes through several different segments and the overall score

obtained by averaging the calculations brings a smoother result. However, the differences between these long videos are still clear; for example, “Soccer” has a much higher value of ATI (23.93) than “Tennews” (9.88). Moreover, people’s requirements for comfortable and pleasant viewing varied with content types, evidenced by the interview data analysis (e.g., participants wanted to a higher quality for movies than other contents [41]). Thus, a direct indicator of content type (e.g., CTsport and CTmovie), if available, will be useful for weighting different content types to reflect the users’ needs. Lastly, from the perspective of user study, our study raises the requirement for more high quality benchmarking videos that have a long duration (e.g., 1–5 minutes), covering a variety of content types, and being freely available for research.

VII. Conclusion And Future Work

Based on the acceptability data obtained from user studies on quality acceptance assessment of mobile device, this paper concentrates on the quality acceptability-based QoE (A-QoE) modeling.

Using statistical techniques, we proposed a set of eight mathematical QoE models to predict pleasant acceptability PAcc (i.e., the possibility for regular and comfortable watch) for various usage situations in mobile video applications. The general acceptability GPAcc (i.e., the possibility for acceptable watch) can be measured based on the close cubic relationship between PAcc and GAcc. The developed A-QoE models used the influencing factors of user acceptance as the model predictors and mapped the variation of the acceptability with the changes of its influences. The proposed QoE models can achieve high prediction accuracy (R>0.9, R^2>0.85), and can be applied into the mobile video system to benefit consistent user perception and effective resource allocation.

Investigations were also undertaken on three full-reference objective video quality assessment metrics – PSNR, SSIM, and VQM, in order to examine whether and how the objective measurements of video quality are related to the subjective acceptability. In terms of Spearman Rank Order Correlation (SROC) and Pearson Linear Correlation (PLC) coefficients, the perceptual video quality predictions given by the SSIM and VQM metrics had a close correlation with the subjective quality acceptability (R>0.6). However, their prediction accuracy and monotonicity are far less than the proposed A-QoE models, and they can be improved by involving other influencing factors such as SR, FR and content features into the model.

Concerning the limitations of the user studies: lab context and task scenario-based test process, our future work will focus on developing a A-QoE-based mobile video system, and then conducting an empirical user study under real usage scenarios and contexts for evaluating and improving the A-QoE models, and establishing a QoE management strategy for user-centred video services.

Acknowledgment

We acknowledge the important contributions made by participants in our user studies.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transaction on Image Processing,* vol. 13, pp. 600-612, 2004.

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," presented at the 37th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2003.

[3] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting,* vol. 50, pp. 312-322, Sep. 2004 2004.

[4] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of experience-driven adaptation scheme for video applications over wireless networks," *IET Communications,* vol. 4, pp. 1337-1347, Jul. 2010 2010.

[5] T. Zinner, O. Hohlfeld, O. Abboud, and T. Hossfeld, "Impact of frame rate and resolution on objective QoE metrics," presented at the Second International Workshop on Quality of Multimedia Experience (QoMEX 2010), Trondheim, 2010.

[6] Z. Fan, L. Weisi, C. Zhibo, and N. King Ngi, "Additive Log-Logistic model for networked video quality assessment," *IEEE Transactions on Image Processing,* vol. 22, pp. 1536-1547, 2013.

[7] F. Agboma and A. Liotta, "Quality of experience management in mobile content delivery systems," *Telecommunication Systems,* vol. 49, pp. 85-98, 2012.

[8] R. Schatz, S. Egger, and A. Platzer, "Poor, Good Enough or Even Better? Bridging the gap between acceptability and QoE of mobile broadband data services," presented at the IEEE International Conference on Communications ICC 2011, Kyoto, Japan, 2011.

[9] ITU-T Recommendation P.10/G.100 Amendment 1, "Definition of quality of experience," in *Vocabulary for performance and quality of service*. ITU-T SG 12, 2007.

[10] T. C. M. de Koning, P. Veldhoven, H. Knoche, and R. E. Kooij, "Of MOS and men: bridging the gap between objective and subjective quality measurements in mobile TV," presented at the Multimedia on Mobile Devices 2007, IS&T/SPIE Symposium on Electronic Imaging, San Jose, CA, USA, 2007.

[11] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez, "Predicting quality of experience in multimedia streaming," presented at the 7th International Conference on Advances in Mobile Computing and Multimedia, Kuala Lumpur, Malaysia, 2009.

[12] T. De Pessemier, K. De Moor, A. J. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, L. Martens, and R. Van De Walle, "Exploring the acceptability of the audiovisual quality for a mobile video session based on objectively measured parameters," presented at the 3rd International Workshop on Quality of Multimedia Experience (QoMEX), 2011.

[13] W. Song, D. Tjondronegoro, and M. Docherty, "Exploration and Optimisation of User Experience in Viewing Videos on A Mobile Phone," *International Journal of Software Engineering and Knowledge Engineering,* vol. 8, pp. 1045-1075, 2010.

[14] F. Agboma and A. Liotta, "QoE-aware QoS management," presented at the 6th International Conference on Advances in Mobile Computing and Multimedia Linz, Austria, 2008.

[15] I. n. Ketyko, K. De Moor, W. Joseph, L. Martens, and L. De Marez, "Performing QoE-measurements in an actual 3G network," presented at the 2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Shanghai, China, 2010.

[16] J.-R. Ohm, "Bildsignalverarbeitung fuer Multimedia-Systeme," *Skript,* 1999.

[17] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transaction on Image Processing,* vol. 19, pp. 1427-1441, 2010.

[18] A. Eichhorn and P. Ni, "Pick your layers wisely - A quality assessment of H. 264 Scalable Video Coding for mobile devices," presented at the IEEE International Conference on Communications, Dresden, Germany, 2009.

[19] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing,* vol. 19, pp. 335-350, 2010.

[20] P. Coverdale, S. Moller, A. Raake, and A. Takahashi, "Multimedia Quality Assessment Standards in ITU-T SG12," *Signal Processing Magazine, IEEE,* vol. 28, pp. 91-97, 2011.
[21] ITU T Recommendation G.107, "The E-model, a computational model for use in transmission planning," Telecommunication Standardization Sector of ITU, 2005.
[22] ITU-T Recommendation G.1070, "Opinion model for video-telephony applications," Telecommunication Standardization Sector of ITU, 2007.
[23] M. Ries, O. Nemethova, and M. Rupp, "Video quality estimation for mobile H.264/AVC video streaming," *Journal of Communications,* vol. 3, pp. 41-50, Jan. 2008 2008.
[24] J. D. McCarthy, M. A. Sasse, and D. Miras, "Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video," presented at the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, 2004.
[25] S. Jumisko-Pyykkö, V. K. M. Vadakital, and M. M. Hannuksela, "Acceptance threshold: A bidimensional research method for user-oriented quality evaluation studies,"[Open Access] *International Journal of Digital Multimedia Broadcasting,* vol. 2008, p. 20, 2008.
[26] A. Raake, J. Gustafsson, S. Argyropoulos, M. Garcia, D. Lindegren, G. Heikkila, M. Pettersson, P. List, and B. Feiten, "IP-Based Mobile and Fixed Network Audiovisual Media Services," *Signal Processing Magazine, IEEE,* vol. 28, pp. 68-79, 2011.
[27] Technical University of Munich. [Online]. Available: ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/
[28] Sveriges Television AB (SVT). [Online]. Available: ftp://vqeg.its.bldrdoc.gov/HDTV/SVT_MultiFormat/
[29] C. Carlsson and P. Walden, "Mobile TV-to live or die by content," presented at the 40th Annual Hawaii International Conference on System Sciences (HICSS), 2007.
[30] K. O'Hara, A. S. Mitchell, and A. Vorbau, "Consuming video on mobile devices," presented at the SIGCHI on Human Factors in Computing Systems, San Jose, CA, USA, 2007.
[31] W. Song and D. Tjondronegoro, "A survey on usage of mobile video in Australia," presented at the OZCHI 2010, Brisbane, Australia, 2010.
[32] Blender Foundation. (2008). Available: http://www.bigbuckbunny.org/index.php/download/
[33] R. Christensen, *Log-Linear models and logistic regression*, 2nd ed. New York: Springer-Verlag inc., 1997.
[34] ITU-T P.910 Recommendation, "Subjective video quality assessment methods for multimedia applications," 1999.
[35] N. Cranley, P. Perry, and L. Murphy, "Dynamic content-based adaptation of streamed multimedia," *Journal of Network and Computer Applications,* vol. 30, pp. 983-1006, Dec. 2005 2006.
[36] J. Joskowicz and J. C. L. Ardao, "A parametric model for perceptual video quality estimation," *Telecommunication System,* vol. 46, p. 14, 2010.
[37] P. G. Gottschalk and J. R. Dunn, "The five-parameter logistic: A characterization and comparison with the four-parameter logistic," *Analytical Biochemistry,* vol. 343, pp. 54-65, 2005.
[38] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," 2004.
[39] ITU-R Recommendation BT.1683, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," 2004.
[40] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A Classification, review, and performance comparison," *IEEE Transactions on Broadcasting,* vol. 57, pp. 165-182, 2011.
[41] W. Song, D. Tjondronegoro, and M. Docherty, "Saving bitrate vs. pleasing users: Where is the break-even point in mobile video quality?," presented at the ACM Multimedia 2011, Scottscale, Arizona, USA, 2011.

Wei Song received M.E. in Signal and Information Processing from Taiyuan University of Technology, China, in 2008, and Ph.D. degree in Information Systems from Science and Engineering Faculty, Queensland University of Technology, Australia, in 2012. Her research interests include subjective video quality assessment, quality of experience, video coding and delivery, HCI, mobile multimedia, and mobile innovation for real-time transit information system and personalized services.
Dr. Song has been working as a Postdoctoral Research Fellow and a Research Assistant since 2010, funded by Smart Services CRC and CRC Rail Innovation, Australia. She was working as an Engineer, and a Manager of Network Resources in Datong Branch, China Netcom Group Corp. Ltd. during 1996-2005, where she gained expertise in telecommunication network system analysis, design and management.

Dian W. Tjondronegoro received Bachelor of IT with first class honours in Queensland University of Technology, Australia, in 2002, and PhD degree in Deakin University, Australia, in 2005. He became a Member (M) of IEEE in 2009.
He has been an Associate Professor in Service Sciences Discipline, Faculty of Science and Engineering, Queensland University of Technology, Australia, since 2011. His research interests are: multimedia computing and communication, and mobile/Web interactions and applications. He has published over 95 academic publications in these fields, including IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 40(5), 2010; Journal of the American Society for Information Science and Technology, 60(9), 2010; and ACM Transactions on Multimedia Computing, Communications, and Applications - TOMCCAP, 4(2), 2008.